Multi Classification Model ROOT2AI Technology Private Limited

Data Science,Machine Learning.

Prepared by Mr. Mayank Pagaria

1.Your thoughts on problem or what was your approach to solve the problem:

The problem is the text classification problem, and our goal is to investigate which supervised machine learning methods are best suited to solve it. This is a Multi Classification problem ,where the dataset consists of just two features one is 'Text' and the other one is 'Target' column.The classifier makes the assumption that each new complaint is assigned to one and only one category. As it is text data we have to use from sklearn feature selection use text to get into numerical for that use count_vectorizer and tfidf transformer.

Split the data into X_train, X_text, y_train, y_test in the ratio of train data 75% and test data 25%.

Creating Prediction Model using algorithms of classification model of supervised Learning algorithm and which has the highest accuracy we will use that model to predict our results on test data

2. Model Interpretation:

1. **Importing all necessary Libraries**: Firstly I have Imported all the libraries like,pandas,sklearn,tfidfTransformer, CountVectorizer,matplotlib,Smoke Tomek for sampling.

2. Data Collection: With the help of pandas I have Imported the dataset.Which was given to me.

3. Data Cleaning: I have checked whether there are any null values or not,Then I found there are 3 null rows.So I have filled them space character instead of removing them. There are totally 22701 rows in the dataset,3 rows doesn't matter at all.

4. Data Visualization: To check there are any ImBalanced data or not. I have used Count plot,Which helped me a lot.There I found that the tuples named "FinTech" have more than 1 8000 records which are completely Imbalanced dataset.

5. Converting Text to Numbers: Our Systems can only understand the number format so we need to convert the given text format to numbers. To do this I have used the CountVectorizer technique for Text column and for Target I have used label Encoding and then applied.

6. Convert to Tfldf: Once The OverSampling is done then we can happily convert them into vectors with the help of Tf Idf.(Which will give us the frequency count of the words in a particular sentence).Finally we got our data as a Balanced and in numeric format.

3. Train & test accuracy score:

1. Model Selection: I have used the

RandomForestClassifier: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

LinearSVC: The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. If we compare it with the SVC model, the Linear SVC has additional parameters such as penalty normalization which applies 'L1' or 'L2' and loss function. The kernel method can not be changed in linear SVC, because it is based on the kernel linear method.

MultinomialNB,The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

LogisticRegression : In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc

PassiveAggressiveClassifier : The Passive-Aggressive algorithms are a family of Machine learning algorithms that are not very well known by beginners and even intermediate Machine Learning enthusiasts. However, they can be very useful and efficient for certain applications.

The model selected on the bases of Which is the best one for Textual data applications. And the LinearSVC is best for such Text Classifier.

2. Testing: Please change the ‘test’ list and give your own text and check.It will definitely match with the TARGET CLASS.

3. Accuracy: Once the training is done,The accuracy was around 54% which was better and the results of the model are pretty much working fine. 2

4. Confusion Matrix: A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

5. Hyper parameter Tuning: Once the training is done with Hyperparameter tuning,The accuracy was around 55% which was better and the results of the model are working fine.

4. Limitation of the model. T

The limitations of the model is accuracy score,And testing with other algorithms are left.But the results are working fine.If you want to test please change the test list and fill with your own text.Then execute it. The results for me are satisfying.So please check this once.

It computes document similarity directly in the word-count space, which may be slow for large vocabularies It assumes that the counts of different words provide independent evidence of similarity. It makes no use of semantic similarities between words. TF-IDF is based on the bag-of-words (BoW) model, therefore it does not capture position in text, semantics.

5. **You can add your own points as well This is how I have implemented the multi classification problem**

I Can Increase the accuracy by performing other algorithms like GridSearch CV to Classifie Algorithm . from GridSearch CV We can increase the model Accuracy.s. As LinearSVC is best for classification type data so I think it will be good fit for the model and the model will be generalized with low bias and low variance. The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. If we compare it with the SVC model, the Linear SVC has additional parameters such as penalty normalization which applies 'L1' or 'L2' and loss function. The kernel method can not be changed in linear SVC, because it is based on the kernel linear method.